Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (11): 2107-2110, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2107.2110

Method of Clustering Web Pages Based on Granular Computing

[1]Jun Hu, [2]Chun Guan and [1]Bocheng Liu
[1]School of Software, Nanchang University, Nanchang, Jiangxi, 330047, China
[2]School of Information Engineering, Nanchang University, Nanchang, Jiangxi, China

Abstract: With the fast development of information on the web, relevant web pages can be discovered by analyzing web server log files and user access database. Based on the granular computing theory, a method of clustering web pages is proposed in this study. Firstly, according to the users accessing the website URL, the method can coarse grain the universal set gradually by granular operation and processing similar grains technology. Then, it can adjust the threshold value to make the division of the universal set be a relatively reasonable status. Finally it can implement the optimal adjustment of websites.

Key words: Granular computing, web intelligence, web log mining, clustering, method

INTRODUCTION

Nowadays, Web data mining technology is becoming a hot study point in current computer field (David *et al*., 2013). There are a huge amount of information resources on the web and there are also a large amount of users to use the information. Therefore, many enterprises attract more and more users to pay more attention to their enterprises by setting up their websites so that many users can obtain the relative information of their enterprises. For internet e-commerce, customers' browsing information is collected by Web server automatically and saved in the Web server log. Through the effective analysis on Web log, we can optimize and adjust the topological structure of Web sites as well as make much effective market strategies for enterprises and help enterprises confirm the target market, improve the strategies and win much bigger competitive advantages (Hasuike and Ichimura, 2013; Russell and Yuhui, 2007; Zhang, 2005).

For a specific business website of internet e-commerce, its topological structure is already known. Although different user has different browsing modes at different stages, the trend should be stable for a long term/Therefore, the relative pages of business website could be discovered by analyzing users' access information for a contain period.

Users' access information basically is recorded in Web log, Web server log includes access log, quoted log and agent log. The common web log normally adopts ECLM log mode which mainly these fields that is, IP Address, Auth ID, Time/Data, Method /URL/Protocol, Status, Size, Referer and Agent. Among them, Time/Data stands for the time when the Web server receives the request from users; URL stands for URL address of accessing page requested by users; Referer means URL of referrer (points to the page of requested file). If the user input URL directly to visit or use book mark to visit, then this column is empty. Agent means the type of the browser and operating system used by the user for browsing. After merging and eliminating these logs, we express the Web server logs in this form: L = ⟨ip, uid, url, time⟩. Among them, uid,ip,url,time separately mean User's ID, User's IP address, User's requested URL and its corresponding time. Therefore, we can deal with it furthermore so that it can reflect the browsing behavior for a certain period.

In this study, by analyzing web server log files and customer's transaction data, the author puts forward a method of clustering web pages based on granular computing, According to web site's directed graph defined, a relevant matrix Url-UserID is set up, where Url is taken as a row, UserID is taken as a column and element's value is taken as the user's hits, relevant web pages are obtained by granular computing, With the method, we can effectively adjust the structure of Web pages and add appropriately the relative additional links; therefore, we can increase the page views of users.

GRANULAR COMPUTING

The concept of Granularity was proposed by Hobbs (Hobbs, 1985). The term "Granular Computing" was suggested by Lin (2001) which was approved by Zadeh (1998). Granular Computing (GrC) is a new way to simulate human thinking to help solve complicated problems and involves all the theories, methodologies and techniques of granularity, providing a powerful tool for the solution

Corresponding Author: Jun Hu, School of Software, Nanchang University, Nanchang, Jiangxi, 330047, China

of complex problems, massive data mining and fuzzy information processing. In the past few years, there are a renewed and fast growing interest in GrC. Granular computing has begun to play important roles in bioinformatics, e-Business, security, machine learning, data mining, high-performance computing and wireless mobile computing in terms of efficiency, effectiveness, robustness and uncertainty (Bianchi *et al*., 2013; Ding *et al*., 2013; Han and Lin, 2010).

The study (Yao, 2004a) gave an overview of Granular Computing. In modeling granular computing, the author focused on three basic components and their interactions as follows:

- **Granules:** Granules are regarded as the primitive notion of granular computing. A granule may be interpreted as one of the numerous small particles forming a larger unit. Collectively, they provide a representation of the unit with respect to a particular level of granularity. That is, a granule may be considered as a localized view or a specific aspect of a large unit. The size of a granule is considered as a basic property. Intuitively, the size may be interpreted as the degree of abstraction, concreteness, or detail. In the set-theoretic setting, the size of a granule can be the cardinality of the granule
- **Granulated views and levels:** A level consists of entities called granules whose properties characterize and describe the subject matters of study, such as a real world problem, a theory, a design, a plan, a program, or an information processing system. Granules are formed with respect to a particular degree of granularity or detail. The granularity is reflected by the sizes of all granules involved. A granule in a higher level can be decomposed into many granules in a lower level and conversely many granules in a lower level can be combined into one granule in a higher level
- **Hierarchies:** Granules in different levels are linked by the order relations and operations on granules. The order relation on granules can be extended to granulated views (levels). A level is above another level if each granule in the former level is ordered before a granule in the latter level and each granule in the latter level is ordered after a granule in the former level, under the order relation. The ordering of levels can be described by the notion of hierarchy. A hierarchy represents relationships between different granulated views and explicitly shows the structure of granulation

With the introduction of the three components, one can examine three types of structures for modeling their interactions. They are the internal structure of a granule, the collective structure of the all granules and the overall structure of all levels. The three structures as a whole is referred to as the granular structure.

METHOD OF CLUSTERING WEB PAGES BASED ON GRANULAR COMPUTING

The user's browsing behavior could be expressed in this form: 2×(n+1) tuple: B = ⟨ipB, uidB, {(lB.url,hits)}n⟩, among which, lBϵL, lB.ip = ipB, lB.uid = uidB, n≥1; hits means the times which the customer uidB has browsed the pages lB.url so far.

Web sites could be expressed as the following directed graph: G = (N, Np, E, E p), among which N is a nodal set; Np = {NodeϵN, {(UserID, hits)}n}, n≥1 which records the customer UserID and its times of accessing the node which is a nodal attribute set;E is a directed edge set; Ep = {(eϵE, {Number of path}p)}m, p, m≥1 which records the directed edge and it's code of path, is a directed edge attribute set. We could get all URL on the web site from the nodal set N of the directed graph G,and could obtain the UserID and its access times of every nodal set from the corresponding nodal attribute set Np. Therefore, we can set up the following transposed matrix M_{mHn} of the relevant matrix Url-UserID.:

$$M_{m\times n} = \overbrace{\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,j} & \cdots & h_{1,n} \\ h_{2,1} & h_{2,1} & \cdots & h_{2,j} & \cdots & h_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{i,1} & h_{i,2} & \cdots & h_{i,j} & \cdots & h_{i,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{m,1} & h_{m,2} & \cdots & h_{m,j} & \cdots & h_{m,n} \end{pmatrix}}^{\text{User ID}} \Big\} \text{URL}$$

Among which $h_{i,j}$ is the times for the user j to visit the i URL for a contain period, every row of the vector M[., j] means the access situation of all URL"." for all users. Every column of the Vector M [i, .] means the access situation of all URL"." for the user "." on its business website.

To effectively fulfill the Web page clustering, In this study, based on setting up relevant matrix of Url-UserID, the author puts forward a new clustering method of web pages based on granular computing, According to web site's directed graph defined, a relevant matrix Url-UserID is set up, where Url is taken as a row, UserID is taken as a column and element's value is taken as the user's hits, relevant web pages are obtained by granular computing and the optimal clustering result can be obtained with the concept of similar granulation.

Definition 1: Similar grain. Setting ϕ and ψ are two Rough logic formulation in Information System s randomly, $m(\phi\wedge\psi)$ and $m(\phi\vee\psi)$ are two grains of Information System S, $t = |m(\phi\wedge\psi)|/|m(\phi\vee\psi)|$ is similarity, $m(\phi)$ and $m(\psi)$ are similar grains and if only setting $t>\theta$ (θ is a threshold value, $0\leq\theta\leq1$).

The specific procedures of this method are as follows:

Step 1: On $M_{m\times n}$ of the relevant matrix Url-UserID, Mapping it to a information system S, then, S = (U, A, V, f)

Among them, U=Url,which is a nonempty finite set, is called a individual all-around of Url, A =UserID, could be regarded as a nonempty finite set of an individual attribute and every UserID is an attribute, V = Url-UserID which is the range of A that is:

$$V = \bigcup_{a\in A} V_a = \bigcup_{1\leq j\leq n, 1\leq i\leq m} h_{ij}$$

V_a, is called the value set of $a\in A$), f is a mapping function, f: $U\times A\rightarrow V$, $\forall u\in U$ has $f(u, a)\in V_a$ for $\forall a\in A$

Step 2: For every attribute subset B ($B\subseteq A$) in S, we could define a indistinguishable relationship R on $U\times U$: $R(B) = \{(u, u')\in U\times U: \forall a\in B \text{ has } f(u, a) = f(u', a)\}$

Step 3: According to the indistinguishable R (B) on U, we get the equivalence class of division U/R(B) which recorded as [u]B, [u]B = {u'U: u R(B) u'}, is a grain

Step 4: Making B = A = URL, to divide U is a clustering web page based on granular computing for a universal set,and every equivalence class is a relevant grain of page of the universal set

EXPERIMENTS

Assuming there are 10 URLs in a business website,and there are totally 4 users to visit the website for a period and every Url's situation visited by users are as follows, please see Table 1. Therefore, we could set up the following transposed matrix M10×4 of the relevant matrix UserID-Url (To use 1 to instead of all non-zero numbers in the matrix)

Setting B = T = {u1, u2, u3, u4}, then, there is U/R(B) = {{r1, r6, r9}, {r2}, {r3}, {r4}, {r5, r8}, {r7}, {r10}}, that is, one type of similar clustering web page for a universal set based on granular computing is:to divide the universal set U to 7 grains that is to make URL r1, r6, r9 to be one grain, URL r5, r8 to be one grain and URL r2, r3, r4, r7, r10 to be one grain separately.

Table 1: User access Url list

	UserID			
Url	U1	U2	U3	U4
r1	0	0	23	0
r2	13	0	31	56
r 3	7	0	18	0
r4	0	20	0	0
r5	0	25	9	44
r6	0	0	27	0
r7	15	0	39	67
r8	0	26	7	51
r9	0	0	30	0
r10	0	17	0	23

With this method of graining, we clustered Url pages, e.g., on the page r1, we add additional links to pages r6 and r9 and the recommendation of this typical page inevitably improve the staying and browsing time and satisfaction evaluation and so on. Meanwhile, it still has some defects while users are increasing more and more, the division on the universal set U must become much specific and the amount of related grain of cluster must be bigger and bigger which makes graining lose some concern significance.

Since, the main reason to lead to the bug occurring is that there are too many users, so we should try our best to decrease the amount of users for a comparative reasonable division amount, we shall merge some similar users. But how to judge two users are similar, the author is introducing relative concept of grain.

In $M_{10\times4}$ of the relevant matrix UserID-Url, setting ϕ = (u2, 1), ψ =(u4, 1), $m(\phi\wedge\psi) = \{r5, r7, r8, rl0\}$, $m(\phi\vee\psi) = \{r2, r4, r5, r7, r8, r10\}$, getting the threshold value $\theta = 0.5$, then $|m(\phi\wedge\psi)|/|m(\phi\vee\psi) = 4/6>0.5$, then $m(\phi)$ and $m(\psi)$ are similar grains. Therefore, there is, u2 and u4 are similar users and we can merge the second and fourth column of the matrix to be user u24 that is, to cooperate some relevant columns, e.g., setting u2 as 1, or u4 as 1, then, u24 value is 1, if the:

$$M_{10\times4} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

values of u2 and u4 are 0, then the value u24 is 0. For the details, please refer to Table 2.

Setting B = {u1, u3, u24} repeatedly, then, there is U/R(B) = {{r1, r6, r9}, {r2, r7}, {r3}, {r4, r10}, {r5, r8}} that

Table 2: Simple list for User access Url

	UserID		
Url	U1	U3	U24
r1	0	1	0
r2	1	1	1
r 3	1	1	0
r4	0	0	1
r5	0	1	1
r6	0	1	0
r7	1	1	1
r8	0	1	1
r9	0	1	0
r10	0	0	1

is one graining strategy based on granular computing for a universal set is: To divide U into 5 grains, comparing to the situation before emerging, the division on the universal set is rougher and the amount of the grains is smaller, the complexity of the system is lower and the amount of similar page grains divided by the universal set U is much practical in application.

CONCLUSION

In this study, the author puts forward a method of clustering web pages based on granular computing. With this method, we could coarse grain the universal set gradually by granular operation and processing similar grains technology according to the users accessing the website URL and we also can adjust the threshold value to make the division of the universal set be a relatively reasonable status and finally implement the optimal adjustment of websites. In addition, through processing graining computing of transposed matrix $M_{m\times n}$ of the relevant matrix Url-useID, we could discover similar customer groups. Since, the processing process is similar to the clustering web pages, in this study the author will not state more. And about how to put granular computing technology into the relative Web content study effectively and reasonably, is a question that has practical application value and we shall study it further.

ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China (Grant No. 11226042) and the Science Research Foundation of Jiangxi Educational Committee (Grant No. GJJ12050).

REFERENCES

Bianchi, F.M., L. Livi, A. Rizzi and A. Sadeghian, 2013. A granular computing approach to the design of optimized graph classification systems. Soft Comput., 10.1007/s00500-013-1065-z

David, C., R.D. Maria and A. Rajendra, 2013. Challenges and issues of web intelligence research. Proceedings of the 3rd International Conference on Web Intelligence Mining and Semantics, Jun 12-14, 2013, New York.

Ding, S., H. Huang, J. Yu and H. Zhao, 2013. Research on the hybrid models of granular computing and support vector machine. Artificial Intell. Rev., 10.1007/s10462-013-9393-z

Han, J. and T.Y. Lin, 2010. Granular computing: Models and applications. Int. J. Intell. Syst., 25: 111-117.

Hasuike, T. and T. Ichimura, 2013. Web intelligence for tourism using railway data by a simplified fuzzy reasoning method. J. Intell. Fuzzy Syst., 24: 251-259.

Hobbs, J.R., 1985. Granularity. Proceedings of the 9th International Joint Conference on Artificial Intelligence, August 18-23, 1985, Los Angeles, USA., pp: 432-435.

Lin, T.Y., 2001. Granular fuzzy sets: A view form rough set and probability theories. Int. J. Fuzzy Syst., 3: 373-381.

Russell, C.E. and S. Yuhui, 2007. Computational Intelligence. Morgan Kaufmann Publishers, USA.

Yao, Y.Y., 2004. Granular Computing. Proceedings of the 4th Chinese National Conference on Rough Sets and Soft Computing Computer Science, Volume 31, June 2004, USA., pp: 1-5.

Zadeh, L.A., 1998. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information intelligent systems. Soft Comput., 2: 23-25.

Zhang, Y.Q., 2005. Computational Web Intelligence and granular Web intelligence for Web uncer-tainty. Proceedings of the IEEE International Conference onGranular Computing, Volume 1, July 25-27, 2005, Atlanta, GA., USA., pp: 99-101.